DOCUMENT RESUME

ED 194 571 TM 800 700

AUTHOR Holmes, Barbara J.
TITLE Bias: Psychometric and Social Implications for the National Assessment of Educational Progress.
INSTITUTION Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO NAEP-12-IP-54
PUB DATE Aug 80
GRANT NIE-G-80-0003
NOTE 37p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Assessment: Educational Objectives: Federal Programs: *National Surveys: *Social Bias: *Statistical Bias: *Test Bias: Test Construction: Testing Problems: Test Items
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

In recent years, the controversy surrounding testing has grown, and the charge of bias is the most often cited criticism of testing and assessment. A review of the literature indicates that psychometricians and other researchers speak of bias as a property of the test or of items in the test. Conversely, test critics speak of bias as a quality or attitude associated with testing, which is suggestive of cultural preferences. The meaning of bias from the research and social perspectives is explored and the areas of conflict between the two are pointed out. By delineating some of the primary points in these divergent perspectives, the position of the National Assessment of Educational Progress relative to cultural bias is clarified in terms of its unique purposes and design. The National Assessment has implemented several techniques during the processes of objectives and exercises development and administration to minimize the intrusion of cultural bias or preferential material. The process of consensus and the evaluative criteria used to develop, write and select exercises are described. Ways in which a new analysis variable, achievement class, can be utilized as a bias detection method with the National Assessment exercises is discussed. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED194571

BIAS: PSYCHOMETRIC AND SOCIAL IMPLICATIONS FOR THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

No. 12-IP-54

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

T. Pratt

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

by Barbara J. Holmes

Assisted by the Staff of the National Assessment of Educational Progress

Education Commission of the States
Suite 700, 1860 Lincoln Street
Denver, Colorado 80295

August 1980

TM 800 700

The National Assessment of Educational Progress is funded by the National Institute of Education. It is under contract with the Education Commission of the States. It is the policy of the Education Commission of the States to take affirmative action to prevent discrimination in its policies, programs and employment practices.

The National Assessment of Educational Progress is an education research project mandated by Congress to collect and report data, over time, on the performance of young Americans in various learning areas. National Assessment makes available information on assessment procedures and materials to state and local education agencies and others.

The work upon which this publication is based was performed pursuant to Grant No. NIE-G-80-0003 of the National Institute of Education. It does not, however, necessarily reflect the views of that agency.

TABLE OF CONTENTS

PREFACE 1

INTRODUCTION 3

EXPLORING THE MEANING OF BIAS 7
Bias as a Psychometric Property 7
Bias as a Social Quality 9
Conflicts Between Psychometric and Social Perspectives of Bias 11

MINIMIZING CULTURAL BIAS IN THE DEVELOPMENT AND ADMINISTRATION OF NAEP EXERCISES 17
Objectives Development 19
Exercise Development 20
Administration 21

CONCERNS AND PROPOSED FUTURE DIRECTIONS 25

BIBLIOGRAPHY 31

NOV 17 1980

PREFACE

The following paper is one of seven required by the National Institute of Education (NIE) grant to the Education Commission of the States for the administration of the National Assessment of Educational Progress (NAEP). The papers are intended to raise significant questions about NAEP functions and procedures in order to apprise the Assessment Policy Committee of its technical and administrative options as NAEP moves into a second decade of activity. According to the NIE Program Announcement, each paper should include "discussions about the significance of a national assessment of educational progress for (a) informing public opinion, (b) influencing educational and other social policies, (c) identifying specific educational and research needs and (d) contributing to the goal of improving education." Thus, the papers combine broad theoretical considerations with very concrete discussions of potential options and their trade offs. All the papers are suggestive, rather than definitive -- starting points for discussion rather than end points.

Two fundamental questions undergird these papers. First, if one were to design a national assessment program from scratch to serve the needs of the eighties and nineties, what would it look like? What theory would it draw upon? What constituents would it serve? How would it operate? And second, given that an assessment has been operating for 10 years, how do answers to these questions match the current NAEP? How could the current NAEP be changed to accommodate needs and technology unforseen when it was begun in the 1960s? As these questions weave in and out of the discussions of various issues, it will become clear that the answers one gives will probably depend upon one's assumptions about such things as the way society, education and science work, as well as one's familiarity with the theory, politics, history and operation of the current National Assessment.

There are many ways to implement the 1978 legislation mandating a national assessment. Some are undoubtedly better than others in certain respects; all have both advantages and drawbacks, complicating even the

simplest comparisons among alternatives. In such a situation, it is imperative that advocates for various models of a national assessment adopt an open, interdisciplinary approach to rethinking the Assessment, listening carefully to each other and respecting professional differences of opinion.

Individuals charged with making policy decisions about NAEP -- members of the Assessment Policy Committee and members of the U.S. Congress -- will find that, in many instances, attempts to strengthen the Assessment in one regard will weaken it in some other regard. Given a certain level of funding, advances in some areas come at the expense of others. The National Assessment can neither be all things to all people nor a single, unchanging enterprise. It will always represent a combination of various tensions. The question for its future is not how do we eliminate those tensions? Rather it is: How do we harness those tensions creatively to keep the Assessment abreast of the times and continually relevant to the needs of American education?

INTRODUCTION

This paper examines cultural bias within the context of a large-scale assessment program, the National Assessment of Educational Progress (NAEP). Bias, as it pertains to testing, is an old and complex issue. Alfred Binet and Theodore Simon acknowledged the possibility of cultural bias in 1905 after developing the first test of intelligence. In 1911, Binet attempted to minimize the effect of cultural bias in that test by eliminating those items that appeared to be overly dependent on specific information of the type most likely attained in school or in an educated home.

Sensitivity to this issue did not fade with Binet's death. In 1940, Raymond Cattell introduced the concept of "culture free" tests, later changing this to "culture fair," recognizing that nothing produced by humans can in fact be culture free.

In 1945 Kenneth Eells and Allison Davis performed an item analysis of 650 items selected from a cross-section of standardized tests administered at that time to several ethnic groups in America. In 1951, Eells et al. published Intelligence and Cultural Differences, now regarded as a classic work in this area. Because Eells did not find conclusive evidence of cultural bias, he cautioned:

> The presence of such a large proportion of unexplained differences should lead to caution in accepting the idea that all status differences on test items can be readily accounted for in terms of the cultural bias of their content (Eells et al., 1951, p. 357).

Nevertheless, Eells and Davis continued to support the ideal of constructing culturally fair tests.

After a hiatus, concern about cultural bias in tests was reawakened as a part of the "heredity versus environment" controversy. In 1969, A.R. Jensen published his now famous article suggesting that intelligence is influenced more by genetic than by environmental factors (Jensen, 1969, pp. 1-123). This article gen-

erated wide discussion reminiscent of that heard during the eugenics movement of the early 20th century. Unfortunately, although Jensen's conclusions about the source of differences in intelligence between blacks and whites were refuted immediately by many scholars (Kagan, 1969; Hunt, 1969; Elkind, 1969; Brazziel, 1969), the refutations have not received equal media attention and have thus failed to be as influential in the public mind as the Jensen article.

Opponents of the genetic-intelligence theory view testing as the most recent, socially sanctioned attempt to perpetuate institutional racism by assuring the maintenance of a racist-inspired class system. Stated differently, testing is perceived by some as a tool used to justify various forms of educational inequity which, in turn, can limit access to the full range of economic and social opportunities in this country. R.L. Williams (1971, 1972, 1974), L.J. Kamin (1974) and Hilliard (1977, 1979a) are representative of many scholars opposed to the genetic-intelligence theory and to widespread intelligence testing.

As test critics charge bias, psychometricians have devised and refined ways of detecting certain kinds of test bias. T.A. Cleary (1968), R.L. Thorndike (1971), D.M. Medley and T.J. Quirk (1974), W.H. Angoff (1972, 1973, 1975) and L.M. Rudner (1977, 1979, 1980) are but a few of the scholars who have investigated and compared means of detecting bias in tests. However, this corpus of research has tended to obscure the major thrust of criticism by focusing on mean score, standard deviations and correlation coefficients rather than on some of the realities of daily life as perceived by test takers and test critics.

Because there is some tension between the positions of test critics and psychometricians about the presence and effect of cultural bias in testing, this paper will first synthesize some of the literature representative of the social and psychometric perspectives, and point out the conflicts between them.

The second section of this paper will describe objectives and exercise development and administration in terms of the safeguards employed by the National Assessment of Educational Progress (NAEP) to minimize the effect of cultural bias in assessment exercises (items). The final section of the paper will discuss some of the questions left unresolved by the vast body of literature on testing and cultural bias and suggest several ways in

which National Assessment data can have heuristic value as answers are sought for certain of these questions.

EXPLORING THE MEANING OF BIAS

Bias refers to an intangible -- it is not concrete in the way an object is concrete or tangible. In the research world, bias refers to an internal property of a test. This property pertains to the responses elicited by the test itself and can be identified through statistical inference. In the social world, bias refers to a quality that motivates and perpetuates a set of values. In turn, this quality may be associated with the test externally and internally. For instance, the content of the test (internal) may reflect a single cultural interest. On the other hand, factors surrounding a test (external), such as various environmental situations, motivation of test takers, expectations of testers and/or use of test results, may be imbued with the quality of bias. Given this general frame of reference, bias assumes different meanings and characteristics in the psychometric and social worlds.

Bias as a Psychometric Property

Educational and psychological researchers assume that certain mental abilities are universal in the human species and that these can be identified through means of various tests. A test score becomes a measurement of the degree to which the individual possesses a particular mental ability. Some researchers (Jensen and Jencks are contemporary examples) assume that detecting variation in mental abilities across groups of individuals assists the educational process. Thus, a test score is presumed to be indicative of the cognitive development or readiness required to master or to perform certain educational tasks. Such scores are, in turn, used for selection, placement and evaluation of individuals. The test score is the basis for an inference.

Within this frame of reference, bias is conceived as a property of the test or of certain items in the test. It means that a test or a group of items in the test do not behave the same way when administered to two or more groups of individuals (respondents). Stated differently, a biased test does not elicit responses that are different in degree (which would suggest a nat-

ural and expected variation among the responses), but rather elicits responses that are different in kind (which would suggest variations due to measurement of somewhat different traits). If the latter occurs, the researcher, in effect, cannot attain a valid or accurate measure of the mental ability that is being investigated. Therefore, the test is discriminating among two or more groups of respondents on the basis of some factor other than that which the test purports to measure.

If a norm- or criterion-referenced test is biased, its predictive value is reduced. This problem is pointed out in T.A. Cleary's (1968, p. 115) definition of bias. Cleary said ". . . the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup," and emphasized that ". . . there may be a connotation of 'unfair' . . . if the use of the test produced a prediction that is too low."

Because of the potentially deleterious effect of bias, many psychometricians have researched methods for its detection and have made extensive comparisons between various bias detection methods. According to L.M. Rudner and his associates (1979, pp. 1-2), there are two major types of investigations of psychometric bias: the first type is concerned with the fairness of the instrument as a whole and concerns whether a test unduly favors or impedes respondents from different parts of the country or from different backgrounds. Methods for detecting test bias have been proposed by Cleary (1968), Thorndike (1971), Darlington (1971), Cole (1973), Einhorn and Bass (1971), Faggen-Steckler et al. (1974) and Gross and Su (1975).

The other major type of investigation concerns the items within a test and asks which items and item formats are appropriate for a given population and which may be used across all specified groups. Rudner indicates that the item bias investigation is of concern to test developers "as it assists them in developing valid and cross-culture fair items while providing a framework for constructing better tests in subsequent efforts." Research of the latter type is also extensive and is represented in the work of Angoff (1972), Cardall and Coffman (1964), Green and Draper (1972), Merz (1973, 1976), Ironson and Subkoviak (1979) and Rudner et al. (1979, 1980).

Item bias, as opposed to test bias, is of concern to the National Assessment and will be discussed in the

final section of this paper.

Bias as a Social Quality

The concepts of culture, racial group and ethnic group are the source of some misunderstanding which, in turn, obscures the meaning of cultural bias. Confusion about the meaning of culture is of no surprise since social scientists themselves use more than 60 definitions of culture (Kroeber, 1952). This confusion carries over to our understanding of racial and ethnic groups since these concepts hold no significance except in the context of culture.

These concepts emerged from extensive anthropological and sociological research with discrete populations over the past 80+ years. In the context of much of this research, a population group shares a historic origin, a symbolic code, geographic and political boundaries and more often than not, a gene pool as well. Thus, such groups can be defined as racial and ethnic groups and their interaction over time results in a particular culture.

Consider a broad definition of culture that focuses on the result or outcome of a group's common experience over time:

> Culture refers to all the socially learned and transmitted patterns of behavior, mutual expectations, common understandings and values an individual shares with others of his group (Anthropology Today, 1971, p. 544).

Because American society is pluralistic, intergroup distinctions become nebulous. America contains several racial groups and many ethnic groups. Some racial groups do not share a similar historic origin, although, in the past, they shared a gene pool. An example of this can be found among Caucasoid Americans, many of whom do not share historic origins. Moreover, Caucasoids are merged into various ethnic groups, who, in the past, shared a similar historic origin. An example of this can be found among people classified as "Hispanos." This major American ethnic group is comprised of major racial groups: Caucasoids, Negroids and Mongoloids who, over time, have shared an historic environmental and political situation.

Pierre L. Van den Berghe (1970, p. 10) defined "racial group" in terms particularly characteristic of intergroup relations in America:

> A racial group is a group of people who in a given society are socially defined as different from other groups by virtue of certain real or putative physical differences. What makes a society multiracial is not the presence of physical differences between groups, but the attribution of social significance to such physical differences as may exist.

While the notion of physical differences is subsumed in the definition of racial groups, in the definition below of ethnic groups, the central point is what people think and believe about themselves in terms of referent group identity.

> An ethnic group refers to people who conceive of themselves and are regarded by others as belonging together by virtue of common ancestry, real or fictive, and a common cultural background (Anthropology Today, 1971, p. 545).

People can thus perceive themselves and can be perceived by others as belonging to the same ethnic group irrespective of their race. The concept of ethnicity, in fact, is perhaps more powerful than that of race in today's world because most races, particularly those that have been thrown into contact over time in America, are no longer genotypically discrete, although phenotypes may still be distinctive.

While American society is comprised of a heterogeneous population, this fact, in and of itself, would not prevent the development of a distinctive American culture. In fact, an American culture has developed from the potpourri of races and ethnic groups within its common geographic boundaries, although the values of that culture have been dominated by the group in largest number: white Americans of Western European descent.

Many of our educational and social problems stem from the tradition of classifying people by race and/or ethnicity. Moreover, the imposition of the values and expectations of a single group on the fabric of American culture and society has led smaller groups to zealously guard and maintain certain distinctions -- languages, dialects, customs and rituals, religious beliefs and

other tangible manifestations of their ethnic identity. The dynamics of ethnicity are consequently reflected in testing just as in all other aspects of American life.

Within the context of testing, cultural bias refers to the degree to which a test or a group of test items discriminates between groups of respondents on the basis of intercultural knowledge. In other words, cultural bias (when bias is conceived as a test artifact) exists when a test measures an ability largely dependent upon the respondent's or group's identification with the symbols and values expressed implicitly or explicitly in the content of the test. When a test is dominated by the values of one particular culture group variance in scores or group results between two or more groups are in fact due to variance in identification with a set of cultural characteristics rather than to variance in the mental ability under investigation.

Conflicts Between Psychometric and Social Perspectives of Bias

Within a social context, bias has a different connotation than it has in the world of psychometrics. Increasingly, the word has come to be synonymous with prejudice, especially racial prejudice. Denotatively, the two words are closely related, since they refer to a preference, but a bias may be objective or subjective in nature, whereas a prejudice does not rest on objective fact. Prejudice is an attitude adopted without respect to information or empirical knowledge.

An outcome of the social dynamics of the 1960s has been a shift in meaning of several words and the cessation of usage of some words. Bias and prejudice are examples of shifts in meaning since bias has come to replace prejudice in popular, social usage.

Given these variations of meaning, the arguments of many test critics are based on a different set of assumptions than those of psychometricians. Test critics generally do not assume that the full spectrum of human mental ability can be measured, nor do they assume that it is essential to identify levels or degrees of mental ability in order to educate people. Consequently, many of their criticisms revolve around broader definitions of bias.

First, for instance, critics claim bias in the <u>practice</u> of testing -- whether intelligence or achieve-

ment testing. In this context, bias refers to a cultural value supposition about the efficacy of testing. This claim is illustrated in remarks by Maria Ramirez, New York State Department of Education, at the National Conference on Achievement Tests and Basic Skills (March 1978). She asserted that there is no such thing as culture-free achievement testing inasmuch as the very act of achievement testing assumes a cultural value placed on achievement.

> Some groups are more cooperative than aggressive, more deferential than competitive. Rewards are primarily for the group not for the individual. Thus, the first cultural bias of achievement testing is the very concept itself (Ramirez, 1979, p. 8).

Ramirez concluded by suggesting that the question becomes not how to eliminate cultural bias from achievement tests, but how to be sure that the biases are valid. Stated differently, bias in tests should be declared and should be defensible. For example, if an achievement test containing items derived from content materials not associated with school curriculum is administered to children in school, it becomes a biased test. However, if the purpose of the test is to discover what children know about that content in order to develop a grade or difficulty level for new curriculum or to serve other diagnostic applications, this is a defensible bias and it can be declared with impunity. In this instance, the bias is consistent with the purpose of the test.

Second, critics claim bias in the items used in tests because content is often not cross-culturally relevant. Asa G. Hilliard (1979b) suggests that, while systematic assessment of achievement is important, certain assumptions and constructs that undergird widespread testing and standardization should be examined from the point of view of cultural variation. Hilliard asserts,

> . . . the cultural bias of a test cannot be corrected simply by changing the number of African-Americans in the standardization sample, as was done on the revised Wechsler Intelligence Scale for Children. The problem of cultural bias is neither in counting nor in the sample, but in the items. The items must be culturally appropriate (Hilliard, 1979b, p. 215).

Cultural appropriateness has to do with the respondents' identification and familiarity with the language, concepts and values contained in the tests. For example, the language used to present the items can be primarily characteristic of one cultural group or ethnic group. Test items used as a basis for deducing cognitive skills may reflect a preference for a particular cognitive style. Test items may express values by introducing objects and other tangible artifacts associated with a particular group.

A study undertaken by Medley and Quirk (1974, pp. 235-245) demonstrated the interaction between item content and race of respondents. Medley and Quirk applied a factorial design to responses to general culture items on the National Teacher Examination and found that cultural bias in item content accounted for significant variations in performance between white and black teachers on the examination.

Jones (1975, pp. 30-33), Padilla and Garza (1975, pp. 53-58) and Taylor (1975, pp. 72-80) have made similar observations about item content in intelligence and achievement tests. Jensen notes this, too, but offers no objective ways of assessing cultural loading.

Third, critics claim bias in the use of test results. The concerns of many are reflected in these words by W. Findley (1973, p. 38):

> When most people think of using tests to group school children for instruction, they naturally think of classifying children of the same grade into different classroom groups, often a high, a middle and a low group. The expectation is that the children will learn more if taught with others of similar general ability . . . Experience and analytical studies have shown that ability grouping fails to make good on its promise to produce better learning.

In connection with the use of test results, W.W. Turnbull (1979, p. 6), president of the Educational Testing Service, identified three fallacies inherent in the excessive expectations of test users.

> The "micrometer fallacy" attributes to test scores a precision -- "an infallibility" -- that they have never possessed. The "whole person fallacy" tends to cause much more to be

> read into achievement test scores than the amount a student has learned in a given subject. The "equal preparation fallacy" leads to beliefs that the test compensates for the differences in academic development of children whose learning opportunities have differed dramatically.

B.C. Watson (1979, pp. 7-8), Temple University, has succinctly described the overreliance on numbers that stems from excessive expectation of test users.

> We believe that if we can take human qualities or characteristics, reduce them to quantitative terms, analyze them with sophisticated techniques, and come out with a correlation for a series of them . . . we've got it! . . . It's a convenience, but it certainly does not represent reality.

Test critics agree that some assessment of achievement is important, but insist that the limitations of testing be recognized and that tests be appropriate to the cultural groups to whom they are administered. Psychometricians have countered the claims of critics by engaging in increased exploration and refined applications of statistical procedures to identify any statistical bias that may be present in widely administered standardized tests of intelligence and achievement. The most recent example is presented in Jensen's book Bias in Mental Testing. He concludes, after a comprehensive statistical examination of the most widely used standardized tests, that (1) such tests measure what they purport to measure and (2) such tests are not psychometrically biased because they reliably predict future education attainments for diverse groups. In other words, the tests are reliable cross-culturally in terms of what they measure (Jensen, 1980, p. 737). It is also noteworthy that while defending standardized intelligence and achievement tests, Jensen concedes that widespread standardized intelligence testing is unnecessary to the act of providing education to all citizens in the country (ibid., p. 716). He recommends, instead, tests of achievement, both norm- and criterion-referenced, that are directly based on the instructional programs offered in schools.

However, Jensen and other psychometricians negate certain fundamental problems about testing. While a test may behave the same way in cross-cultural administrations, the assumptions underlying testing, the con-

tent of the test and the way test results are used may still constitute biases. Psychometric defenses of tests do not negate the value of raising some fundamental questions about equity in a democratic society as ways to improve testing are sought.

MINIMIZING CULTURAL BIAS IN THE DEVELOPMENT AND ADMINISTRATION OF NAEP EXERCISES

Cultural bias in testing remains a complex issue -- an issue still requiring consistent monitoring and additional study to assure that all populations within America are fairly tested and assessed. Definitional prob - lems have in part contributed to the complexity of the issue. Complexity also emerges from the practice of separating specific manifestations of cultural bias from the overall rubric of cultural bias. Often researchers and critics speak of specifics such as regionalisms, dialectal variation, sex roles and status as something apart from culture bias. These specific types of bias occur within a larger cultural frame of reference so an attempt to separate them from the culture concept leads to some misapprehension of the point. For the remainder of this paper, cultural bias will include sex, race, region, language, economic and role and/or status preferences.

In this discussion of cultural bias, preference is the operable word. A cultural bias will be said to exist whenever a preference for a particular sex, race, region, language, role or status is exhibited in exercises although that preference in no way contributes to the accuracy or precision of content as it relates to objectives the exercises are designed to measure. Preferential material differs from a general cultural perspective in exercises. For example, assessment exercises developed by American educators for administration to American students will no doubt have a cultural perspective distinctive from that of assessment exercises developed by Chinese educators for administration to Chinese students. However, assessment exercises, containing content based on an actual body of knowledge in a subject area, perceived by qualified educators as germane to the acquisition of knowledge and understanding in that subject area, should not be construed as culturally or psychometrically biased simply because they elicit differential responses across population groups. Rather, it is the nonessential accentuation of preferential material that moves toward deleterious cultural bias. Bias in testing is far more complex than mere differences in responses.

Before considering some of the cultural bias concerns in developing and administering assessment exercises, it is useful to examine the purposes of the National Assessment and some of the distinctions between achievement and aptitude or ability testing.

The purpose of the National Assessment is to collect, measure and report data on the achievement -- not aptitude -- of American students. The National Assessment is a continuing, census-like survey or indicator of achievement among population groups both at one point in time and over time. Assessment results do not permit a rank-ordering of individuals, nor do they provide a basis for prediction. However, Assessment results can be used to do several things:

- Identify gaps in achievement between groups (implications for education policy makers)
- Identify changes in achievement between and within groups over time (implications for education policy makers and practitioners)
- Diagnose patterns of achievement between and within groups (implications for education practitioners and designers of intervention strategies)
- Generate hypotheses about factors associated with gaps in achievement between and within groups (implications for a variety of secondary researchers)

As D.R. Green (1976) pointed out, an assessment of achievement means that the amount of knowledge students have acquired at certain points in time is measured. Differences in amounts learned at certain points in time are to be interpreted as resulting from the adequacy of instruction, materials, prior achievement and motivation, as well as students' aptitude. Conversely, differences in aptitude measures or scores are usually interpreted as resulting from variations in aptitude. So the purpose of the National Assessment sets it somewhat apart from tests such as the Wechsler Intelligence Scales (for adults and children), the Stanford-Binet Intelligence Scale, the Scholastic Aptitude Test, the School and College Ability Test and other tests of this type. Each of these tests is norm- and/or criterion-referenced, and each is designed to have some predictive applications.

The National Assessment shares with other test developers sensitivity to culturally preferential material that can obscure the measurement of achievement. NAEP's chief means of safeguarding the validity of a general cultural perspective are the consensus and review processes that drive the various phases of an assessment cycle. These processes and other steps will be discussed below in terms of minimizing cultural bias in objectives development, exercise development and administration. For a more thorough discussion of objectives and exercise development, the reader should see The National Assessment Approach ... (1980).

Objectives Development

One way to discuss issues of concern to NAEP is to make a distinction between external and internal sources of cultural bias. For example, external sources may be environmental situations, the people and attitudes that surround the development and administration processes. Another external source of bias may be the way assessment or test results are used. Internal cultural bias pertains to the actual exercises -- their content, format, level of difficulty and the relevance of distractors, and so forth. In short, internal cultural bias is a property of the exercises.

National Assessment tries to minimize external cultural bias by developing objectives with the following considerations in mind:

- Objectives are developed to maximize a student's chance to exhibit knowledge acquired from a wide range of nonschool sources as well as from in-school exposure and instruction.

- Objectives are developed to reflect the diversity in each assessed subject area. The subject area should not be defined in narrow or selective terms, excluding the curricular goals of many American schools; rather, they should reflect the expansions occurring in each assessed subject area.

- Objectives are developed to represent the fullest possible range of behaviors or skills considered important in the subject area, including those that are difficult to measure via pencil and paper exercises.

The National Assessment involves people from all ethnic and social groups, from all regions of the country, and of both sexes in the task of developing assessment objectives. These persons have experience at several professional levels within the subject area (e.g., academicians and theoreticians working in university and other research settings, classroom teachers, curriculum specialists and concerned lay persons). They negotiate objectives until they reach consensus. Consensus does not mean that the majority viewpoint drives out the minority view; nor does it mean that everyone agrees upon bland, "lowest common denominator" objectives. Rather, it means that, through negotiation, objectives that are important to any group will find their way into the final statement of goals. It is an inclusive, rather than an exclusive consensus process. In this way, certain perspectives or preferences that could result in a cultural bias are minimized because the objectives do not represent a single curriculum or emphasis in the subject area.

Exercise Development

National Assessment exercises are used in a slightly different manner than exercises designed for norm-referenced and criterion-referenced tests. Assessment exercises are designed to discover how defined groups of students -- as opposed to individual students -- respond to content in a given subject area. NAEP exercises do not allow for a single, cumulative-type score for individuals or groups. Rather, the Assessment collects and reports data on the percentage of students able to answer a single exercise or a set of related exercises (related by objective and/or content).

Because exercises are of paramount importance to the credibility and utility of the National Assessment program, several safeguards are implemented to minimize cultural bias in them. A series of interrelated steps defines the exercise development process. Generally, these steps are: writing and review, rewrite and review, tryouts in the field, review and possible additional rewrite, and final selection.

A process of negotiated consensus drives exercise development just as it does objectives development. Persons engaged in the process must be skilled exercise writers as well as knowledgeable about the subject area. As in objectives development, persons writing and re-

viewing exercises are involved at various levels of the education system and represent the ethnic, regional and social diversity of the nation. Persons performing exercise reviews are not the same as those who write exercises, but are equally qualified in the subject area.

Panels of exercise writers and reviewers are guided by the following evaluative criteria:

- Exercises should measure objectives.
- Exercises should cover a broad spectrum of subject-area content.
- Exercises should range in complexity from "easy" to "difficult" and should not include irrelevant distractors.
- Exercise formats should be mixed, that is, not all of a multiple-choice format.

In addition to these criteria, reviewers confirm accuracy of exercise keys, precision of language, face validity of the exercises, and then review the total exercise set to assure a balance between objectives and the content.

In recent years, the National Assessment has expanded its contacts with minority professional education organizations. As a result, participation in objectives and exercise development by minority educators and lay persons has increased. However, the Assessment staff's concern has not ceased. Efforts to expand minority participation are reviewed and modified periodically to assure that minorities are involved in all subject-area assessments and that all minority groups are represented in the task.

Administration

To assure uniform administration of exercises and to minimize the burden for participating schools, a professional data collection staff from the Research Triangle Institute administers each assessment. The data collection staff is trained in appropriate administration procedures to ensure maximum control in all national assessments. The use of a trained staff of outsiders not only assures standardized procedures, but also reduces the effect of some external factors associated with testing, such as teacher/student familiarity, teacher

expectations and teacher favoritism.

Administration is standardized in order to minimize the effects of external factors that could hamper or alter student performance:

- Each age group is assessed at approximately the same time of the school year so that students in one year's assessment are at about the same place in curriculum from one assessment to the next.

- Exercise booklets are administered to small groups of students in their schools and during the normal school day. The administration of the exercise booklet requires only slightly more time than a typical class period (40-45 minutes). Several administration groups may be working on assessment exercises during the same school day and at the same time.

- Exercises are administered to students with a paced audio-tape to compensate for any variations in reading skills and to control timing.

- Administrators may stop the session if there are external factors in the physical environment that may hamper or impede students as they work on the exercises. Examples of such factors are undue noise from outside the administration room, an emergency causing noise or other interference from another part of the school, and so on.

The National Assessment has been concerned for some time about relying on the administrator's ethnic identification of students. Prior to 1977-78, administrators coded students' ethnic classification on the exercise booklet. In addition to the administrator's visual identification, 17-year-olds were asked to provide their ethnic identification in the 1977-78 mathematics assessment and in the 1979-80 reading assessment. Consequently, the National Assessment staff was able to compare student's self-identification with the administrator's visual identification. Results of this study indicated high reliability between the two ethnic identifications, with only 5% of the 17-year-olds not responding to the question. In future assessments 13-year-olds, as well as 17-year-olds, will be asked to make an ethnic self-identification, but administrators will continue to be the sole source of ethnic identifi-

cation for 9-year-olds.

Until a baseline can be established for measuring changes in group performance over time, both administrator's visual identification and student self-identification will continue.

In addition to modification of the procedure for obtaining ethnic classifications for students, the ethnic composition of the field administration staff is periodically examined to assure that significant numbers of minority persons are involved as field administrators and district supervisors. Some progress has been made in this area, but improvement is still possible.

Generally, administration procedures are rigorous, but the National Assessment will continue its active monitoring and reevaluations of this major aspect of the Assessment program.

CONCERNS AND PROPOSED FUTURE DIRECTIONS

As the National Assessment selects a course for the future, it is wise to delineate those things the Assessment can and cannot do, given its primary purpose and design features. While NAEP is sensitive to the concerns of test critics and desires to be responsive to the needs of the wider education community as well, some of these concerns cannot be directly addressed through means of the Assessment's broadly descriptive data. Moreover, many of the concerns of test critics are not yet answerable through means of other forms of education data either.

Many of the questions raised in the research literature on education and testing do not lend themselves to easy answers; and even if some answers are forthcoming, there will be a time lag between discovery and incorporation into the education mainstream. For example, the issue of genetic versus environmental influences on human intelligence remains unresolved in spite of Jensen's comprehensive attempt to justify the soundness of currently administered intelligence tests. A study by Sandra Scarr and Richard Weinberg (1976) strongly indicates that intelligence (if one accepts intelligence as a measurable construct) is significantly influenced by environmental factors. These findings counter Jensen's implicit conclusion, and it is interesting to note that Jensen did not cite this study.

Another issue indirectly related to the question of intelligence, with implications for education processes, is that of the relationship between language acquisition and motor skill development in early childhood. Also related is the theory of critical stages of brain development in early childhood. This research (see Lenneberg, 1967, 1973; Luria, 1974; Darley, 1967; Brazier, 1975; and Desmedt, 1977) strongly suggests that if certain forms of environmental stimulation are absent in early childhood, environmental enrichment at a later stage will have limited effect because a critical stage has been passed. In other words, the effects of certain kinds of environmental deprivation are irreversible. Within the context of this research, the effect of spe-

cific kinds of deprivation (nutrition, isolation, and so forth) are not yet clearly understood, especially in terms of their potential impact on the early childhood of those in minority groups.

Although NAEP is greatly interested in the direction of this research and its potential importance to educational process, baseline data of the type collected and reported by the National Assessment cannot be used to throw much light on these issues. However, the National Assessment proposes to elaborate certain of its analyses and reporting strategies to increase the utility of its data for policy makers, education practitioners and other researchers.

The Assessment Policy Committee (the governing body of the National Assessment) has encouraged staff to review and reevaluate the reporting categories (defined population groups) traditionally used as a means of presenting NAEP data. The decision to report data by defined population groups was made during the formative years (1963-69) of the Assessment by members of the Technical Advisory Committee, now the Analysis Advisory Committee. Race and a series of socioeconomic variables were included in the NAEP methodology in order to be responsive to the federal government's need for descriptive data on education achievement among groups of students. It was believed that identifying differentials in achievement between groups could be helpful in ameliorating education inequity in America. However, after a decade of data collection and reporting, the Assessment Policy Committee and the staff are not certain whether this type of achievement data has been utilized in positive or negative ways.

As a result of several meetings of the Assessment Policy Committee and the Analysis Advisory Committee, the racial/ethnic reporting category and some others have been reviewed. The outcome of this review, which included exploration of alternative analyses and reporting strategies, is the development of a variable that will allow data to be presented with greater explanatory power. The new variable is called "achievement class."

Briefly, this variable partitions the sample of respondents by knowledge levels (six have been defined). Many new and informative analyses, within the existing capacity of data processing technology, can be conducted by crossing achievement class with one or more other variables, such as community type, region, race, level of parental education, and so on. As a result of the

achievement class variable, a variety of questions can be addressed, which in turn will reveal more information about the nature of student achievement in America. For example:

1. Measurement of Change

   Does change in performance over time occur uniformly or differentially within each achievement class? For example, given an average three percent change in p-value from one assessment year to the next assessment year (for a group or for the nation), is there also about a three percent change in each of the p-values for the achievement classes or is there a greater change in the lower achievement classes?

2. Exploration of the Characteristics of Low and High Achievers

   NAEP's current approach focuses on the achievement levels of various reporting groups, but does not answer questions about the demographic characteristics of low or high achievers.

3. Detection of Bias

   Is an exercise (or subset of exercises) biased against a particular population group in the sample? By comparing the mean p-values for various categories of a variable, the effect of ability on performance on the exercise(s) is largely removed. The mean package scores on exercises should be close to each other for any group within each achievement class. This is the case regardless of demographic characteristics of those in the achievement class. If the p-values for each exercise by any group variable are largely the same within each achievement class, there is no discernible bias in the exercise(s). On the other hand, if the p-values for some of the exercises are lower or higher within an achievement class for a particular group, then those exercises may be biased for or against that group.

If exercises behave differently within an achievement class, there are several steps the National Assessment can take to discover the underlying reasons for a particular pattern of responses. The exercises can be examined for obvious problems such as a misprint or typographical error in the text of the exercise, in the distractors or in graphics accompanying the exercise. Exercise reviews following tryouts often eliminate prob-

lems of this type and are a standard part of the NAEP procedure. Another aspect of this procedure includes examining exercises against the following criteria:

a. Language in the exercise -- single words or the entire message (tone, values, and so on) conveyed by the exercise

b. Accuracy of the exercise key or correct foil and relevance of other distractors

c. Reading level assumed by the exercise -- complexity of sentence structure relative to the content of the exercise

d. Format of the exercise -- horizontal v. vertical presentation of text and distractors, more v. fewer words, graphics v. no graphics

e. Specific content matter in the exercise -- does it measure the subject-area objective or subobjective? Is the content germane to the subject-area curriculum?

Although the staff is alert to many of these problems in exercises, an ethnically diverse panel of consultants can be invited to review exercises, particularly those exhibiting differential patterns of response. If exercises meet the criteria stated above, they need not be removed from an assessment since they are consistent with the purpose of the National Assessment. However, if exercises meet criterion "e," but do not meet criterion "a" through "d," they may be modified to alleviate the problem. But, if criterion "e" is not met after modification of any of the other four points, the exercise should be removed. Stated differently, criterion "e" is essential to any exercise in an assessment.

Still another aspect of exercise review is the nature of subject area itself. For example, exercises in a subject area such as citizenship/social studies or literature may have more than one correct answer, depending on a student's cultural perspective. Such items can have dual keys.

When exercises meet the criteria discussed above, another level of information is available even though they may behave differently across population groups. Once exercises are determined to be accurate and justifiable in terms of the subject area, detected item bias

can be a means of identifying subtleties of student performance. In turn, this kind of information can be used diagnostically. For example, information about item difficulty levels can be indicative of patterns of student misapprehension of the item content. Janice Scheuneman (1979) examined a group of biased items from the 1976 revision of the Metropolitan Readiness Tests (MRT) after using the chi-square method for item bias detection. The identified items revealed a pattern of negative sentence structure, which appeared to be biased against black students. As a result of using an item bias detection method, followed up by examination of the item content, the test developers were able to make an informed decision about retaining or discarding the items consistent with the objectives and purpose of that particular test. This kind of procedure can be implemented with National Assessment exercises because of its large sample.

Various diagnoses are possible and feasible with NAEP data. For instance, responses to all exercise distractors can be examined for possible patterns of incorrect answers. This could provide insight about the effect of the number of words in the distractors, vertical versus horizontal layout of the distractors, and so on, and the selection of a particular distractor by population groups.

A similar type of diagnosis possible with NAEP data concerns the achievement of population groups relative to the objectives and content developed for a subject area. Following the application of an item bias detection method, important information about student achievement in terms of specific learning objectives and content is available. A simplified analysis of black student achievement in science and mathematics has been performed with National Assessment data and indicated specific areas where early intervention might be helpful for black students (Holmes, 1979). This information, in turn, has assisted the Committee on Minorities in Engineering as they plan and develop intervention strategies.

These are but a few of the issues that can be addressed by using item bias detection techniques with NAEP data. However, due to limitations of resources, the National Assessment staff will not be able to pursue every feasible question that may be forthcoming from the achievement-class variable and item bias detection. NAEP welcomes linkages with other researchers and audiences needing specific kinds of information so that

this capability may be exploited to its fullest. In this way, the National Assessment can continue to serve the changing needs of the wider education community in America.

BIBLIOGRAPHY

Angoff, W.H. "A Technique for the Investigation of Cultural Differences." Paper presented at the Annual Meeting of the American Psychological Association, Honolulu, May 1972.

Angoff, W.H. "The Investigation of Test Bias in the Absence of an Outside Criterion." Paper presented at the National Institute of Education Conference on Test Bias, 1975.

Angoff, W.H. and S.F. Ford. "Item-Race Interaction on a Test of Scholastic Aptitude," Journal of Educational Measurement, vol. 10, 1973, pp. 95-105.

Anthropology Today, Glossary of Anthropological Terms. Del Mar, Calif.: CRM Books, 1971.

Binet, A. and T. Simon. "Methodes Nouvelles pour le Diagnostic du Niveau Intellectual des Anormaux," L'Annee Psychologique, vol. 11, 1905, pp. 191-244.

Brazier, M.A.B., ed. Growth and Development of the Brain: Nutritional, Genetic and Environmental Factors. International Brain Research Organization Monograph Series, vol. 1. New York: Raven Press, 1975.

Brazziel, W.F. "A Letter From the South," Harvard Educational Review, vol. 39, 1969, pp. 348-356.

Cardall, C. and W.R. Coffman. "A Method for Comparing Performance of Different Groups on the Items in a Test." RM64-61. Princeton, N.J.: Educational Testing Service, 1964.

Cattell, R.B. "A Culture-Free Intelligence Test, Part I," Journal of Educational Psychology, vol. 31, 1940, pp. 161-179.

Cleary, T.A. "Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges," Journal of Educational Measurement, vol. 5, 1968, pp. 115-124.

Cole, N.S. "Bias in Selection," Journal of Educational Measurement, vol. 10, 1973, pp. 237-255.

Darley, F.L., ed. Brain Mechanism Underlying Speech and Language. New York: Grune and Stratton, 1967.

Darlington, R.B. "Another Look at 'Cultural Fairness,'" Journal of Educational Measurement, vol. 8, 1971, pp. 71-82.

Desmedt, J., ed. Language and Hemispheric Specialization in Man: Cerebral Event-Related Potentials. Basel S. Karger, 1977.

Eells, K. et al. Intelligence and Cultural Differences. Chicago, Ill.: University of Chicago Press, 1951.

Einhorn, H.J. and A.R. Bass. "Methodological Considerations Relevant to Discrimination in Employment Testing," Psychological Bulletin, vol. 75, 1971, pp. 261-269.

Elkind, D. "Piagetian and Psychometric Conceptions of Intelligence," Harvard Educational Review, vol. 39, 1969, pp. 319-337.

Faggen-Steckler, J. et al. "A Quantitative Method for Measuring Sex 'Bias' in Standardized Tests," Journal of Educational Measurement, vol. 11, 1974, pp. 151-161.

Findley, W. "How Ability Grouping Fails," Inequality in Education, vol. 14, 1973, pp. 38-40.

Green, D.R. "Reducing Bias in Achievement Tests." Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1976.

Green, D.R. and J.F. Draper. Exploratory Studies of Bias in Achievement Tests. Monterey: CTB/McGraw-Hill, 1972.

Gross, A.L. and W. Su. "Defining a 'Fair' or 'Unbiased' Selection Model: A Question of Utilities," Journal of Applied Psychology, vol. 60, 1975, pp. 345-351.

Hilliard, A.G. III. "Standardization and Cultural Bias as Impediments to the Scientific Study of Intelligence," Journal of Research and Development in Education, vol. 12, 1979, pp. 47-58.

_____. "Standardized Testing and African-Americans: Building Assessor Competence in Systematic Assessment," Testing, Teaching and Learning. Washington, D.C.: National Institute of Education, 1979, pp. 204-218.

_____. "The Predictive Validity of Norm-Referenced Standardized Tests," Negro Educational Review, vol. 28 (3 and 4), 1977, pp. 189-201.

Holmes, B.J. "Black Student Performance in the National Assessments of Science and Mathematics: A Diagnostic View." Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, November 1979.

Hunt, J. McV. "Has Compensatory Education Failed? Has It Been Attempted?" Harvard Educational Review, vol. 39, 1969, pp. 278-300.

Hunter, J.E. "A Critical Analysis of the Use of Item Means and Item-Test Correlations To Determine the Presence or Absence of Content Bias in Achievement Test Items." Paper presented at the National Institute of Education Conference on Test Bias, Maryland, December 1975.

Hunter, J.E. and F.L. Schmidt. "A Critical Analysis of the Statistical and Ethical Implications of Various Definitions of 'Test Bias,'" Psychological Bulletin, vol. 83, 1976, pp. 1053-1071.

Ironson, G.H. and M.J. Subkoviak. "A Comparison of Several Methods of Assessing Item Bias," Journal of Educational Measurement, vol. 16, 1979, pp. 209-225.

Jencks, C. et al. Inequality: A Reassessment of the Effect of Family and Schooling in America. New York: Harper and Row, 1972.

Jensen, A.R. Bias in Mental Testing. New York: The Free Press, 1980.

_____. "How Much Can We Boost IQ and Scholastic Achievement?" Harvard Educational Review, vol. 39, 1969, pp. 1-123.

Jones, C. "What Do IQ Tests Measure?" National Elementary Principal, vol. 54, 1975, pp. 30-33.

Kagan, J.S. "Inadequate Evidence and Illogical Conclusions," Harvard Educational Review, vol. 39, 1969, pp. 274-277.

Kamin, L.J. The Science and Politics of IQ. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1974.

Kroeber, A. and C. Kluckhohn. Culture: A Critical Review of Concepts and Definitions. New York: Random House, 1952.

Lenneberg, E.H. Biological Foundations of Language. New York: John Wiley and Sons, Inc., 1967.

_____. "The Neurology of Language," Daedalus, vol. 102, 1973, pp. 115-134.

Luria, A.R. "Language and Brain," Brain Language, vol. 1, 1974, pp. 1-14.

Medley, D.M. and T.J. Quirk. "The Application of a Factorial Design to the Study of Cultural Bias in General Culture Items on the National Teacher Examination," Journal of Educational Measurement, vol. 11, 1974, pp. 235-245.

Merz, W.R. "Estimating Bias in Test Items Utilizing Principal Components Analysis and the General Linear Solution." Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1976.

_____. "Factor Analysis as a Technique in Analyzing Item Bias." Paper presented at the Annual Meeting of the California Research Association, Los Angeles, 1973.

Padilla, A.M. and B. Garza. "IQ Tests: A Case of Cultural Myopia," National Elementary Principal, vol. 54, 1975, pp. 53-58.

Ramirez, M. Comments on "Achievement Testing: An Overview of the Issues," at the National Conference on Achievement Tests and Basic Skills, March 1978. Proceedings published in Testing, Teaching and Learning. Washington, D.C.: National Institute of Education, 1979.

Rudner, L.M. "An Approach to Biased Item Identification Using Latent Trait Measurement Theory." Paper pre-

sented at the Annual Meeting of the American Educational Research Association, New York, April 1977.

Rudner, L.M. et al. "A Monte Carlo Comparison of Seven Biased Item Detection Techniques," Journal of Educational Measurement, vol. 17, 1980, pp. 1-10.

_____. "Biased Item Detection Techniques." Washington, D.C.: National Institute of Education, 1979.

Scarr, S. and R. Wineberg. "IQ Test Performance of Black Children Adopted by White Families," American Psychologist, vol. 31, 1976, pp. 726-739.

Scheuneman, J. "A Method of Assessing Bias in Test Items," Journal of Educational Measurement, vol. 16, 1979, pp. 143-152.

Taylor, E.F. "Science Tests," National Elementary Principal, vol. 54, 1975, pp. 72-80.

Testing, Teaching and Learning. Report of a Conference on Research in Testing. Washington, D.C.: National Institute of Education, August 1979.

The National Assessment Approach to Objectives and Exercise Development, no. 12-IP-55. Denver, Colo.: National Assessment of Educational Progress, Education Commission of the States, 1980.

Thorndike, R.L. "Concepts of Culture Fairness," Journal of Educational Measurement, vol. 8, 1971, pp. 63-70.

Turnbull, W.W. Comments on "Achievement Testing: An Overview of the Issues," at the National Conference on Achievement Tests and Basic Skills, March 1978. Proceedings published in Testing, Teaching and Learning. Washington, D.C.: National Institute of Education, 1979.

Van den Berghe, P.L. Race and Ethnicity: Essays in Comparative Sociology. New York: Basic Books, Inc., 1970.

Watson, B.C. Comments on "Achievement Testing: An Overview of the Issues," at the National Conference on Achievement Tests and Basic Skills, March 1978. Proceedings published in Testing, Teaching and Learning. Washington, D.C.: National Institute of Education, 1979.

Williams, R.L. "Abuses and Misuses in Testing Black Children," Counseling Psychologist, vol. 2, 1971, pp. 62-77.

_____. "Scientific Racism and IQ: The Silent Mugging of the Black Community," Psychology Today, May 1974, pp. 32-41.

_____. "The BITCH (Black Intelligence Test of Cultural Homogeneity) - 100: A Culture-Specific Test." Paper presented at the annual convention of the American Psychological Association, Honolulu, September 1972.